EXHIBIT 2

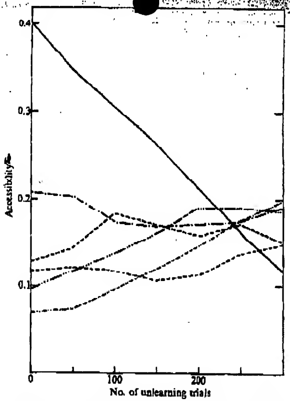

Fig. 1 The fraction of random starting states which leads to particle memories (accessibility). The five dashed lines are the five nominal memories. The solid line is the total accessibility of all spurious memories. In these trials ε was set at 0.01.

'Unlearning' has a stabilizing effect in collective memories

J. J. Hopfield*‡, D. I. Feinstein* & R. G. Palmer†

* California Institute of Technology, Pasadena, California 91125, USA
† Duke University, Durham, North Carolina 27706, USA
‡ Bell Laboratories, Murray Hill, New Jersey 07974, USA

Crick and Mitchison[1] have presented a hypothesis for the functional role of dream sleep involving an 'unlearning' process. We have independently carried out mathematical and computer modelling of learning and 'unlearning' in a collective neural network of 30–1,000 neurones. The model network has a content-addressable memory or 'associative memory' which allows it to learn and store many memories. A particular memory can be evoked in its entirety when the network is stimulated by any adequate-sized subpart of the information of that memory[2]. But different memories of the same size are not equally easy to recall. Also, when memories are learned, spurious memories are also created and can also be evoked. Applying an 'unlearning' process, similar to the learning processes but with a reversed sign and starting from a noise input, enhances the performance of the network in accessing real memories and in minimizing spurious ones. Although our model was not motivated by higher nervous function, our system displays behaviours which are strikingly parallel to those needed for the hypothesized role of 'unlearning' in rapid eye movement (REM) sleep.

In the most symmetric form of collective memory in our dynamic neural network[2], each neurone, j, has two states, and is described by a variable $\mu_j = \pm 1$. The instantaneous state of the system of N neurones can be thought of as an N-dimensional vector having components μ_i of size 1. The neurones are interconnected by a network of synapses, with a synaptic strength T_{ij} from neurone j to neurone i. The instantaneous input to neurone i is

$$\text{input to } i = \sum_{j=1} T_{ij}\mu_j$$

where μ_j is the present state (± 1) of neurone j. The neural state of the system changes in time under the following algorithm. Each neurone i interrogates itself at random in time, but at a mean rate W, and readjusts its state, setting $\mu_i = \pm 1$ according to whether the input to i at that moment is greater or less than zero. The neurones act asynchronously.

This algorithm defines the time evolution of the state of the system. For any symmetric connection matrix, there are stable states of the network of neurones, in which each neurone is either 'on' and has an input ≥ 0 or 'off' and has an input < 0. These stable states will not change in time. Starting from any arbitrary initial state, the system reaches a stable state and ceases to evolve in a time of $\sim 3/W$.

The stable states of the system can be arbitrarily assigned by an appropriate choice of T_{ij}. Suppose n different N dimensional state vectors

$$\{\mu_i^s\}_{i=1}^N,\ s = 1 \text{ to } n \leq 0.25N$$

are to be stable states of the system. If these state vectors are sufficiently different, and if the synaptic connection matrix T_{ij} is given by

$$T_{ij} = \sum_s \mu_i^s \mu_j^s;\ T_{ii} = 0$$
$$i \neq j$$

then the states μ^s will be stable states of the system.

This network now functions as an associative memory. If started from an initial state which resembles somewhat state μ^t and which resembles other $\mu^s (s \neq t)$ very little, the state will evolve to the state μ^t. The states μ^s are evokable memories, and the system correctly reconstructs an entire memory from any initial partial information, as long as the partial information was sufficient to identify a single memory. Detailed properties of the collective operation of this network have been described previously[2].

The form of the T_{ij} matrix can be described as an incremental learning rule. To learn a new memory μ^{new}, increment T_{ij} by

$$\text{learn } \mu^{\text{new}}\ \Delta T_{ij} = \mu_i^{\text{new}} \mu_j^{\text{new}}$$

0028-0836/83/280158—02$01.00

In biology or in circuits, this would be done by placing the system in state μ^{new}—for example, [illegible]o by external inputs—and enabling a learning process that [illegible]ows all T_{ij} to increment. The information needed by each synapse is local—the increment for synapse ij does not depend on the global structure of the new state or past memories, but only on μ_i^{new} and μ_j^{new}.

Under this algorithm, when random starting states are chosen, some stored memories are much more accessible than others, that is, considerably larger numbers of randomly chosen initial states lead to some memories than to others. This is a vagary of the particular set of memories which have been learned. It occurred to us that it would be possible to reduce this unevenness of access (which can be intuitively described as the "50% of all stimuli remind me of sex" problem) by 'unlearning'.

Specific unlearning was implemented by choosing starting states at random; when a final equilibrium state μ^f was reached it was weakly unlearned by the incremental change

$$\text{unlearn } \mu^f \; \Delta T_{ij} = -\varepsilon \mu_i^f \mu_j^f, \; 0 < \varepsilon \ll 1$$

Figure 1 illustrates the effect of unlearning on the accessibility of five stored memories in a set of 32 neurones. Accessibility is quantitatively defined as the fraction of random initial states leading to a particular final stable state or group of states. The unevenness of the lines is due in part to statistical noise in the simulation. The accessibility of the nominal assigned memories initially ranges over a factor of 3, but converges with unlearning to a spread of only a factor of 1.4. Thus the accessibility is much more uniform (or in Crick–Mitchison terms, the relative stability of the modes made more uniform) after specific unlearning, and the system will have functionally improved recall.

In our model the storage of a set of assigned memories in T_{ij} also produces a set of spurious stable states which were not inserted as memory states. One of the strong effects of unremembering is to reduce the total accessibility of spurious states, as shown by the solid line in Fig. 1.

The qualitative reason for the success of unlearning comes from the behaviour of the 'energy' E, defined for any state μ as

$$E = -\sum_{i \neq j} \sum_{j} T_{ij} \mu_i \mu_j$$

The change of neural state with time according to the asynchronous algorithm monotonically decreases E until a final stable state is reached—either a stored memory or a spurious memory. Any stable state μ^m has, for a given T_{ij}, an energy E^m. There is a strong tendency for the states having the deepest energy valleys to collect from the largest number of random starting states, that is, deep valleys are also broad. When a final state μ^f is unlearned, its energy E^f is specifically raised and its valley of collection diminished relative to other states. While this argument indicates why accessibility of stored memories should be made more nearly even by unlearning, only a detailed analysis shows why the spurious states should be so sensitive to it. Too much unlearning will ultimately destroy the stored memories.

We have identified a class of spurious states, which in their most elementary form have their origin in triples. As an example on 16 neurones

Memory 1	+ + + + − − − − | + + − + − + − −
Memory 2	+ + + + − − − − | − − + − + − + +
Memory 3	+ + − − + + − − | + − − + + − − +
Spurious memory	+ + + + − − − − | + − − + + − − +

The stability of the spurious memory is enhanced if the first half of memory 3 is weakly correlated with memories 1 and 2. Mathematical analysis of the statistical stability of such spurious states shows that they are typically less stable than the assigned memories, and that the stability will also depend on correlations with other memories. The nature of these spurious states can be described by analogy in terms of higher level function by the example [illegible]

Memory 1	[illegible]alter, white
Memory 2	Walter, black
Memory 3	Harold, grey
Spurious memory	Walter, grey

where grey is taken as a category equally resembling black and white. This spurious state is more stable when 'Harold' and 'Walter' have a significant correlation—perhaps 'Harold' and 'Harry'. These particular spurious states are not simply transitive logical associations of the form A↔B, B↔C; → A↔C. They are truly spurious 'illogical' associations, but perhaps 'plausible' as they come from correlations in the structure of memories.

In our simple system, unlearning improves memory function both by the equalization of accessibility and the suppression of spurious memories. We asked whether other simple algorithmic changes such as clipping the T_{ij} matrix or a threshold effect produce an equivalent improvement in memory performance. These two do not, presumably because they lack the essential element of the present scheme, that is, the feedback via the algorithm of information about the accessibility of particular states. We believe the results found will be insensitive to whether the state component values are taken as 0 and 1 or ±1.

The REM sleep hypothesis of Crick and Mitchison[1] refers to higher level processing. Our example illustrates that from a mathematical viewpoint the general idea could work as they described. If the Crick–Mitchison hypothesis is correct, one might ask about correlations between the structure of the spurious linkages in modelling and the strange associations present in dreams.

We thank F. Crick and D. Willshaw for discussions. This work was supported in part by NSF grant DMR-8107494 and by the System Development Foundation.

Received 31 December 1982; accepted 15 May 1983.

1. Crick, F. C. & Mitchison, G. *Nature* 304, 111–114 (1983).
2. Hopfield, J. J. *Proc. natn. Acad. Sci. U.S.A.* 79, 2554–2558 (1982).

A language-specific comprehension strategy

Anne Cutler*, Jacques Mehler†, Dennis Norris* & Juan Segui‡

* MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK
† Laboratoire de Psychologie, CNRS, 54 Boulevard Raspail, 75006 Paris, France
‡ Laboratoire de Psychologie Expérimentale, associé au CNRS, 28 rue Serpente, 75006 Paris, France

Infants acquire whatever language is spoken in the environment into which they are born. The mental capability of the newborn child is not biased in any way towards the acquisition of one human language rather than another. Because psychologists who attempt to model the process of language comprehension are interested in the structure of the human mind, rather than in the properties of individual languages, strategies which they incorporate in their models are presumed to be universal, not language-specific. In other words, strategies of comprehension are presumed to be characteristic of the human language processing system, rather than, say, the French, English, or Igbo language processing systems. We report here, however, on a comprehension strategy which appears to be used by native speakers of French but not by native speakers of English.

Underlying our finding is a structural difference between the two languages: French and English differ considerably in the degree to which syllable boundaries are clear and unambiguous. In French, syllabic structure is relatively easily determined[1];